Geographic Information Systems: Their Use in Environmental Epidemiologic Research

Marilyn F. Vine,[1] Darrah Degnan,[1] and Carol Hanchette[2]

[1]Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC 27599-7400 USA; [2]North Carolina Department of Environment, Health and Natural Resources, State Center for Health Statistics, Raleigh, NC 27626-0538 USA

Advances in geographic information system (GIS) technology, developed by geographers, provide new opportunities for environmental epidemiologists to study associations between environmental exposures and the spatial distribution of disease. A GIS is a powerful computer mapping and analysis technology capable of integrating large quantities of geographic (spatial) data as well as linking geographic with nongeographic data (e.g., demographic information, environmental exposure levels). In this paper we provide an overview of some of the capabilities and limitations of GIS technology; we illustrate, through practical examples, the use of several functions of a GIS including automated address matching, distance functions, buffer analysis, spatial query, and polygon overlay; we discuss methods and limitations of address geocoding, often central to the use of a GIS in environmental epidemiologic research; and we suggest ways to facilitate its use in future studies. Collaborative efforts between epidemiologists, biostatisticians, environmental scientists, GIS specialists, and medical geographers are needed to realize the full potential of GIS technology in environmental health research and may lead to innovative solutions to complex questions.

Key words: address geocoding, computer mapping, environment, epidemiology, geographic information system(s), GIS, medical geography, methods, review. *Environ Health Perspect* 105:598–605 (1997)

While the mapping of health data is not new to epidemiologists, advances in geographic information system (GIS) technology provide new opportunities for epidemiologists to study associations between environmental exposures and the spatial distribution of disease. In addition to the conduct of ecologic studies in which environmental exposure information is compared with disease rates across regions at the group level, GIS technology can be used to estimate exposures to individuals in cross-sectional, case–control, and cohort studies. Often the most difficult, costly, and time-consuming aspect of environmental health studies is obtaining accurate exposure information. A GIS can combine information contained in existing databases and/or data that can be computerized to estimate exposure levels, for example, to agricultural pesticides, to individuals residing or working within defined geographic regions. The computerized estimates of exposure, together with information on the location and occurrence of disease among individuals within the regions, can then be used to suggest and support hypotheses regarding environmental causes of disease. So far, only a few studies incorporating GIS technology have been published in the epidemiologic literature. This is partly due to a lack of familiarity with the technology and partly due to limitations in its use for epidemiologic research.

The purpose of this paper is to provide an overview of some of the capabilities and limitations of GIS technology with regard to its use in environmental epidemiologic research; to illustrate, through practical examples, the use of several functions of a GIS including automated address matching, distance functions, buffer analysis, spatial query, and polygon overlay; to discuss methods and limitations of address geocoding, often central to the use of a GIS in environmental epidemiologic research; and to emphasize the need for collaborative efforts between epidemiologists, biostatisticians, environmental scientists, GIS specialists, and medical geographers to realize the full potential of GIS technology in future epidemiologic studies.

What is a Geographic Information System (GIS)?

Essentially, a GIS is a powerful computer mapping and analysis technology that allows large quantities of information to be viewed and analyzed within a geographic context. According to Antenucci et al. (*1*), a GIS ". . . links nongraphic attributes or geographically referenced data with graphic map features to allow a wide range of information processing and display operations as well as map production, analysis and modelling." These techniques allow the health researcher to go beyond the simple mapping of disease rates within predetermined political boundaries (e.g., county, state).

GISs are used to input, store, manage, analyze, and display data. Many GIS experts believe that a true GIS differs from desktop mapping systems in that it contains a data structure that stores information about topology, i.e., the relationships among geographic features (*2*). Certain methods of spatial analysis require a topological data structure, which allows concepts such as adjacency and connectivity, easily visible to humans, to be recognized by a GIS.

Data storage formats. Data can be stored in a GIS two ways: in raster format and in vector format. The raster format stores geographic data or graphic images as a matrix of evenly divided grid cells that contain values for an attribute. The position of the cell in the matrix provides information about location. Additional information about attributes is stored within each grid cell. Raster data can be scanned from maps or obtained from photographs or remote sensing space satellites. Satellite images and digital photos are examples of digital data stored in raster format.

Vector data consist of strings of coordinates and usually are represented in a GIS by three types of features: points, lines, or polygons (areas). A point is represented by a single x,y coordinate in a Cartesian coordinate system that is usually geographically referenced, i.e., tied to real locations on the earth's surface. Lines are typically represented by the x,y coordinates of their beginning and ending points, with intermediate points or vertices defining the shape and curvature of the line. Areas are represented as a boundary made up of a series of connecting line segments.

GIS database development. A GIS database consists of any number of map layers that are referenced to geographic coordinates (e.g., latitude/longitude, State Plane Coordinate System) as well as attributes that can be linked to map layers by a common identifier, or geocode (Fig. 1). Examples of the latter include county-level cancer rates that can be linked to a county boundary map by county codes, demographic characteristics provided by the U.S.

Address correspondence to M. F. Vine, Department of Epidemiology, CB#7400 McGavran-Greenberg Hall, University of North Carolina, Chapel Hill, NC 27599-7400 USA.

The authors would like to thank David Ethridge, formerly of the North Carolina State Center for Health Statistics, for his work in developing Figure 3.

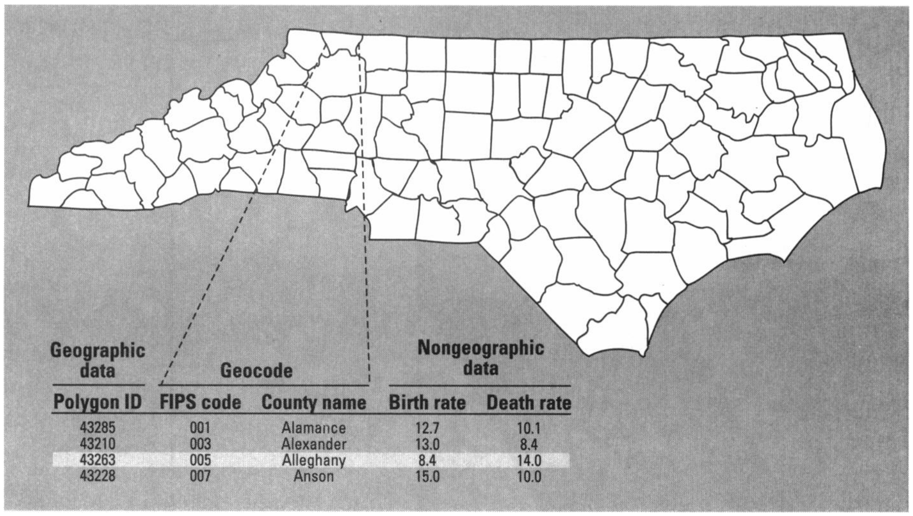
Received 18 November 1996; accepted 14 February 1997.

Bureau of the Census that can be linked to census tracts by geographic codes, or environmental monitoring data that can be linked to specific sites by known geographic coordinates.

The primary bottleneck in the implementation of most GISs is the development of GIS databases (or map layers), which can account for as much as 70% of the time and resources necessary to conduct the research (*3*). Fortunately, many geographic data layers are available through public or private agencies at a reasonable cost. Examples include the U.S. Bureau of the Census TIGER (Topologically Integrated Geographic Encoding and Referencing system) line files, which contain map layers for census geography, street networks, hydrology, railroads, and other man-made features, and the U.S. Environmental Protection Agency's (EPA) Toxic Release Inventory (TRI) Sites file, which provides information on chemicals released into various environmental media (e.g., soil, water).

Although computerized databases may exist, they may not be of the appropriate precision, recency, or completeness necessary to conduct specific research, often having been constructed for other purposes. To be useful in an epidemiologic study, a database must contain spatial coordinates as well as temporal (e.g., dates of exposure assessment) and quantitative information (e.g., level of exposure) regarding the measured factor. After examining 26 environmental databases in California, Frisch et al. (*4*) found that most databases had one or two of these types of information, but few had all three. Similarly, in evaluating the utility of routinely collected health data (such as vaccination information) in England for research purposes, Twigg (*5*) discovered that data were often not of the appropriate spatial detail and that spatial detail varied by source of information.

While the availability of computerized data can reduce the amount of time and money needed for data acquisition, researchers can develop their own databases. Data can be entered into the computer with a digitizer (an instrument that allows a user to trace geographic features with a cursor) or with a scanner. Geographic coordinates for specific locations (e.g., a residence) can also be captured with a handheld global positioning system (GPS) receiver, which interprets signals from three or more satellites that are part of a worldwide positioning and navigation system in orbit 1,500 miles above the earth. Depending on the receiver, the use of and distance from GPS base stations, and the degree of postprocessing, positional accuracy can range from several hundred meters to less than 3 cm (*6*).



Geographic data	Geocode		Nongeographic data	
Polygon ID	FIPS code	County name	Birth rate	Death rate
43285	001	Alamance	12.7	10.1
43210	003	Alexander	13.0	8.4
43263	005	Alleghany	8.4	14.0
43228	007	Anson	15.0	10.0

Figure 1. How a geographic information system (GIS) links geographic and nongeographic data. A polygon ID (from a GIS county boundary file) identifies the highlighted North Carolina county and links it to a file containing a set of geographic coordinates (hidden from the user), which describe the polygon. Geocodes [the Federal Information Processing System (FIPS) and the county name] link the polygon to nongeographic information (e.g., birth or death rates).

GIS database integration. One of the most powerful features of a GIS is the ability to overlay several map layers. When multiple geographic data are stored in a common coordinate system, many map layers can be viewed simultaneously, allowing the user to look through the set of maps in order to understand better the spatial relationships among the features of the different layers. Figure 2 illustrates how a composite map could be created from three overlays, one showing residences and the groundwater wells that serve them, a second showing water-bearing zones, and a third showing those water-bearing zones contaminated from a nearby landfill (*7*). This composite exposure map could be used to assign the residences a likely level of exposure (e.g., high, medium, low) to the contaminated water. The exposure information, combined with knowledge of health/disease status (e.g., immune competence as indicated by immunoglobulin levels) by residential location could be included in a traditional epidemiologic study to investigate associations between exposure to contaminated water and disease risk. Data on potential confounding or modifying factors (e.g., age, race, gender) could also be incorporated into data analyses.

Information about the characteristics of each map layer, often referred to as metadata, is critical when combining layers from various data sources. For example, data sources may not be comparable with respect to 1) the geographic unit to which the data apply (e.g., block, city, county); 2) the scale at which the data were collected (e.g., 1 inch = 2,000 ft or 200,000 ft); 3) the time frame to which the data apply; 4) the accuracy and completeness of the data; or 5) the format in which the data were computerized. Although the final map may look accurate (because it contains clearly recognizable boundaries and landmarks), the combination of incompatible data layers could result in erroneous attribute information (e.g., exposure levels) within the various boundaries (*8*).

Even when data layers are comparable, there are limits to the interpretations one can make regarding overlaid map layers. Often researchers want to make individual level inferences from group level (ecologic) data. The underlying assumption is that the values within map layers reflect the characteristics of the individuals to whom they apply. With regard to exposure levels, however, small pockets of high exposure could be missed if the data were aggregated to a large region. Thus, the map layers might reflect the exposure characteristics of some people but not all, or they might reflect average values.

In ecologic studies, where both the exposure and outcome are measured at the group level, biased interpretations of an exposure/disease association across regions can result at the group level and at the individual level if potential confounders and modifiers of the association are not appropriately taken into consideration in the analyses (*9*). The failure of aggregate level associations to reflect individual level associations is known as ecological bias (*10*). Greenland and Robins (*9*) have presented examples of ways in which bias can occur in ecologic studies, and Morgenstern (*11*) has

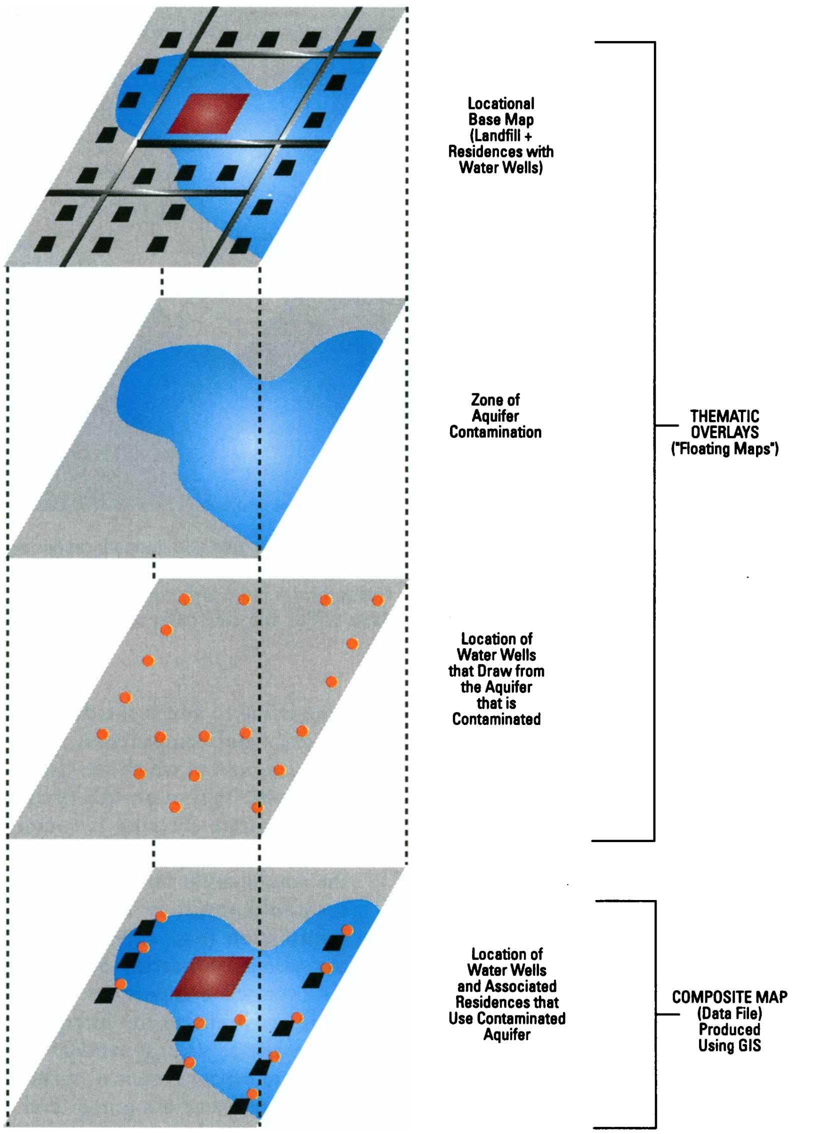

Locational Base Map (Landfill + Residences with Water Wells)
Zone of Aquifer Contamination
Location of Water Wells that Draw from the Aquifer that is Contaminated
THEMATIC OVERLAYS ("Floating Maps")
Location of Water Wells and Associated Residences that Use Contaminated Aquifer
COMPOSITE MAP (Data File) Produced Using GIS

Figure 2. Composite map generated by a geographic information system (GIS) from three map databases. Reproduced from Stallones et al. (*7*) with permission from *Environmental Research.*
Credit: J.R. Nuckols/Carol Wassell

discussed methods of incorporating individual level and group level data into statistical models to try to minimize bias. Biases can also be reduced by analyzing data from smaller geographic units and by stratifying on subgroups with similar disease risk (*12*).

Spatial autocorrelation should be considered in the statistical analysis of spatial data (*13*) because units that are located close together in space tend to influence each other and often have similar characteristics; this violates the assumption of independence in statistical analyses. In fact, a variety of spatial statistical analysis techniques are available, which can be used to control for potential confounding factors and increase the power to detect associations between environmental factors and the spatial distribution of disease (*12,14,15*).

Specialized GIS functions. Despite the above mentioned limitations, GISs possess many features that are particularly useful in environmental epidemiologic research. For example, for data display or map production, most systems provide users with a wide range of mapping options such as colors, symbols, annotation, legends, scales, and other cartographic features as well as the ability to produce charts, graphs, and tables. Other more specialized functions include 1) automated address matching (described in detail below), 2) distance functions (calculation of distances between geographic features), 3) buffer analysis (calculation of a buffer area of a desired width around a point, line, or area), 4) spatial query (the ability to select from a map layer geographic areas with specific characteristics), and 5) polygon overlay analysis (the ability to topologically overlay two or more GIS layers and create a new layer by combining information from the original map layers whose boundaries may not coincide) (Fig. 3).

It should be noted, however, that most GISs have limited statistical capabilities. Output from a GIS is often input into other software for statistical analyses. After the data are statistically modeled, they can be input back into a GIS for display (*8*).

Epidemiologic Studies Using a GIS

The following three examples illustrate the use of several specialized GIS functions for environmental exposure assessment in population studies.

Lead exposure. Guthe et al. (*16*) used existing computerized data to predict populations of children at high risk of lead exposure in the Newark/East Orange/Irvington area of New Jersey. They constructed maps showing census tracts of suspected high lead exposure by overlaying existing databases containing sources of lead exposure (e.g., industrial emissions sources, hazardous waste sites, and traffic volume). Then, they visually compared the census tracts of suspected high lead exposure with census tracts of reported high lead exposure as determined through data obtained from blood lead screening records.

Data from a variety of sources were integrated using Arc/Info GIS software [Environmental Systems Research Institute (ESRI), Inc., Redlands, CA]. Geographic data included 1) census tract boundaries from the U.S. Census Bureau TIGER/Line files, enhanced with data from the ETAK (ETAK, Inc., Menlo Park, CA) database, which contains more accurate boundary information; 2) the locations of lead sources from industrial and hazardous waste sites in the study area obtained from New Jersey Department of Environmental Protection and Energy (NJDEPE) databases; and 3) data concerning vehicle traffic miles/road classifications from the New Jersey Department of Transportation (NJDOT) database and data concerning the locations of roads from the ETAK files. Nongeographic (attribute) data included 1) blood lead screening records from the county health department together with address, sex, date of birth, and date of blood sample, and 2) information concerning populations probably exposed to lead obtained through a spatial query identifying census tracts with ≥620 structures built before 1940 (which would be likely to have lead paint) and with ≥290 children under 5 years of age.

Good, but imperfect, correlations between the census tracts suspected of high lead exposure (based on the data from the existing databases of lead exposure sources and the location of sensitive populations) and the census tracts with reported high lead levels (as determined by the blood lead screening records) motivated investigators to consider additional sources of information to improve the prediction of individuals at high risk of lead exposure. Other possible predictors include lead in drinking water, historic air emissions, soil lead levels, distance of the census tract from a blood screening center, and economic, educational, and cultural factors. Other factors that may have resulted in discrepancies between observed and expected blood lead levels

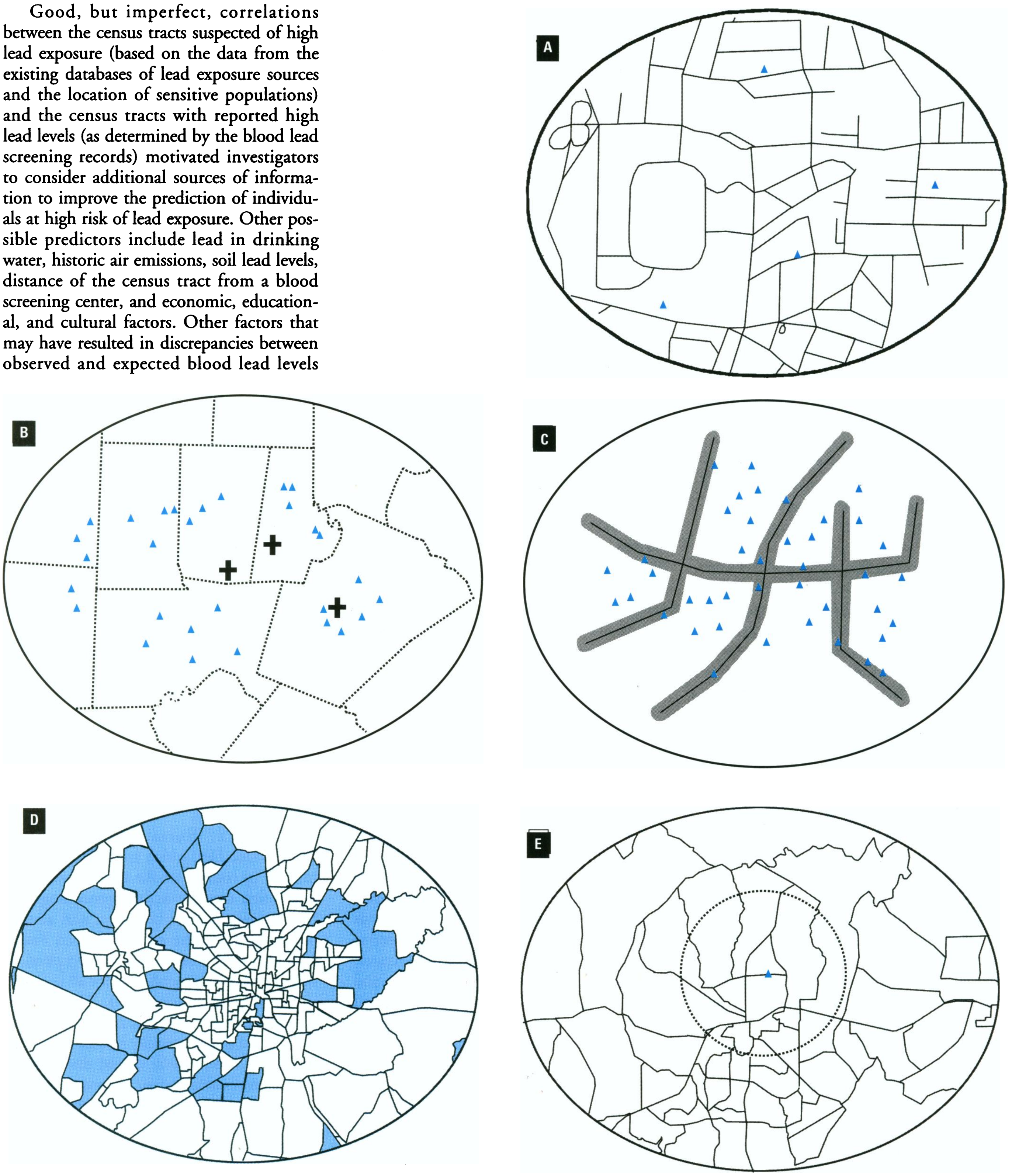


Figure 3. Diagram of five GIS functions. A) Automated address matching can pinpoint the location of a single address, such as 231 Elm Street, by comparing the address to a street network file that contains information about street names and address numbers. B) Distance functions can be used to calculate the distance from study participant residences (triangles) to sources of environmental contamination (crosses) after address matching both. C) Buffer functions can define a geographic area of a desired width around a point, line, or area (for example a 25-m zone around main roads to identify areas with potentially high levels of lead - contaminated soil from past use of leaded gasoline). D) Spatial query allows the user to select from a database geographic regions with specific characteristics (e.g., census block group areas with >150 children under the age of 6). E) Polygon overlay can be used to create a new map layer from map layers whose boundaries do not coincide. For example, one could identify the population within a 1-mile radius of a hazardous waste site (assuming population data were available at the subcensus level) even though the radius does not coincide with census boundaries.

include errors in data entry and reporting of patient addresses, as well as the fact that blood lead levels were only reported if they exceeded the current health standard, which changed many times during the study period (*16*). Furthermore, since the industrial emissions database was constructed in 1987 and the hazardous waste site database was constructed in 1989, it is possible that some sites were not relevant exposure sources at the time the blood samples were taken or that some sites were created after the databases were constructed, and were therefore not included in the study.

A strength of this study was the use of bioassay data (actual blood lead levels) to validate the prediction of an environmental exposure with information in existing databases using GIS methods. Bioassay data are often not available or are expensive and time consuming to collect. If predictors of high lead exposure can be obtained from existing computerized databases or data that can be easily computerized, then predictive models can be constructed in other areas to target high risk individuals for exposure reduction and/or treatment.

In fact, some risk factors for lead exposure can be obtained from the U.S. Census (e.g., residence in older housing, poverty, race/ethnicity) (*17*), allowing screening of high risk communities at the national level. More specific targeting of high risk individuals can be accomplished with the addition of the residential locations of children with high lead levels, if available, and data concerning the locations of other locally relevant environmental sources of exposure.

Electromagnetic field exposure. Wartenberg et al. (*18*) developed a method to identify and characterize populations of sufficient size with potential for high exposure to magnetic fields for epidemiologic cohort studies. They digitized into a GIS each transmission station along a 29-km segment of a 230-kV power line in New Jersey. Using Arc/Info GIS software, they chose a 100-m buffer on each side of the transmission line (corresponding to a field exposure of about 0.2 µT). They used demographic data from the 1990 U.S. Census and TIGER/Line files to map characteristics of the 201 census blocks that were contained wholly or partially within the buffer. Comparison of the demographic characteristics of the people in the buffer area with the people in the surrounding area revealed that the individuals at risk for high exposure were similar to those in the surrounding area, thereby identifying populations suitable for a cohort study assessing the health hazards associated with electromagnetic field exposures.

Using buffers to identify exposed populations works best when a sufficient number of individuals live close to the source of exposure and when the characteristics of the people within the buffer zone are relatively evenly dispersed geographically. However, in this study, for example, for one arbitrarily chosen 230-kV line in New Jersey, the people in the buffer zone who lived closest to the power line were more likely to be white, older than 18, and have less expensive rents than the people in the buffer zone who lived further away. These factors would have to be considered in the analysis. It may be that for other power lines, the people living closest to the power lines were not different from those living further away. It may also be that the composition of the population within the buffer was affected by the inclusion of all people from census blocks that were contained partially as well as entirely within the buffer. Other methods of estimation or interpolation, such as prorating the population by the percent of block group area within the buffer, may have yielded a slightly different population composition.

Environmental risk factors for Lyme disease. Glass et al. (*19*) used a GIS to overlay six different land databases containing 53 environmental variables in order to investigate residential environmental risk factors for Lyme disease in Baltimore County, Maryland. Lyme disease is caused by the bacterium *Borrelia burgdorferi* and transmitted to humans through ticks. High risk areas for Lyme disease include woodland and forest-edge areas. GIS methods were combined with case–control methods to identify residents at high risk of Lyme disease.

The six databases included land use/land cover, forest distribution, soils, elevation, geology, and watersheds. Also entered into the GIS were the addresses of 48 cases of Lyme disease that occurred between 1989 and 1990 and 495 randomly selected control addresses. (Residence was used for the Lyme disease cases because 87% of those infected could identify no other location for infection). The software system used was IDRISI, a spatial analysis software developed by Clark University Graduate School of Geography (Worcester, MA). Residential information for cases and controls was combined with environmental variables in a logistic regression model to determine risk factors for Lyme disease.

A problem with this study and similar studies is that many variables were included in the model relative to the number of cases, decreasing the stability of the model. Even though many variables were considered in the analysis, there could still be unmeasured confounders of the association between location of residence and development of Lyme disease. For example, no individual level information was reportedly obtained from cases and controls with regard to age, sex, or recreational environments and activities.

A strength of the study is that it relied on knowledge of the epidemiology of Lyme disease to identify potential environmental risk factors (e.g., characteristics of the habitats of ticks). Another strength is that it capitalized on the existence of computerized data for the exposure measure but used individual level data for the disease outcome, which permitted individual level statistical analyses. Using existing exposure data also obviated the need to collect expensive survey information concerning tick locations, soil type, land use, etc. If existing computerized data can be useful in predicting Lyme disease cases, these predictors may also be used in other areas, allowing the implementation of preventive measures.

In all of the above examples, one could question the accuracy, completeness, and comparability of the databases that were combined. A critical question is how current were the databases with respect to the outcomes of interest. These factors obviously affect the predictive value of the GIS exposure models. Despite their weaknesses, the above studies represent innovative uses of GIS technology for exposure assessment in environmental health research.

Address Geocoding Function

Address geocoding will be described in detail as it is often a necessary first step in integrating epidemiologic and geographic data. In fact, two of the three above-mentioned examples included address geocoding. Many primary and secondary datasets used in epidemiologic studies contain addresses. Because other geocodes contained in the datasets (e.g., counties or zip codes) may not be specific enough for purposes of the research, residential address geocoding is often used to pinpoint the location of a residence.

The U.S. Bureau of the Census TIGER/Line files, which are defined and distributed by county, are the street network files most often used for address geocoding. The files contain geographic street and address range data for the entire United States (although rural areas are often incomplete), as well as geographic coordinates for all U.S. census units (e.g., block, tract). The smallest address unit within the files is a street segment (i.e., part of a street, usually bounded by intersections of two or more streets) that is coded with a street name and an address range, e.g., 1100–1150. Even and odd addresses usually lie on opposite sides of the street. The location of any particular street address within a street segment is then determined by matching the street name in the address to the same street name in the TIGER/Line files and interpolating within the number range. For reasons discussed below, this method of address geocoding, or address matching as it is often called, is only partially automatic, especially in rural areas. The following example illustrates the benefits and problems of address geocoding in an epidemiologic study conducted in rural North Carolina, as well as other potential uses of a GIS in environmental epidemiologic research.

Perspectives from a North Carolina Study: Actual and Potential Uses of a GIS

The overall goal of a study conducted by the authors in rural North Carolina was to determine whether residents living near several pesticide dump sites were more likely to have evidence of immunosuppression than residents living further away from the sites. A GIS was used to identify the geographic coordinates of study participant residences through address matching and to calculate the distance from each residence to each dump site. Use of a GIS could have facilitated other aspects of the research methods as well.

Selection of study areas. To select boundaries for the exposed and comparison study areas in the county in which the dump sites were located, information concerning residences served by ground versus surface water systems was requested from local water companies. (Study participants were selected from areas served by groundwater systems because groundwater was a suspected route of exposure.) Population data (e.g., racial composition and mean income) were obtained by census block group from the 1990 census. Potential study areas, by block group, were labeled manually on a wall-sized census map with regard to demographic and socioeconomic characteristics so these factors could be considered in the selection of study area boundaries.

Use of a GIS could have facilitated these steps. Census data (demographic and socioeconomic) by block group could have been selected and displayed on the computer providing various options for study area boundaries. Furthermore, an overlay, if available, of the areas served by ground versus surface water, could have been used to confirm that selected areas were on groundwater systems.

Identification of eligible study participants. In order to identify and contact residents in the study areas, a broker list was purchased; this list contained the names, addresses, and telephone numbers of residents living in the selected block groups. The smallest area available for purchase, however, was the zip code. Zip code addresses did not have the same boundaries as the selected block groups, although three zip codes included all of the selected areas. Therefore, each address had to be verified more precisely using local maps and a city directory to determine eligibility in the study. Additional residents and updated telephone numbers were identified with the most recent telephone book. Public water billing records were helpful in updating street addresses, especially in towns without street mail delivery because they listed both the post office (P.O.) box number or rural route number as well as the street address. (P.O. box numbers and rural route numbers are not useful for address matching.)

Residents were asked to confirm their physical street address and indicate the nearest intersecting street during an initial telephone interview.

A GIS could have been used to match residential addresses against the TIGER/Line or other address files to help identify the locations of streets relative to study area boundaries. Of course, the GIS would not have been able to locate P.O. box addresses, but it would have reduced some of the time necessary to identify residential locations.

Address geocoding. A random sample of residents (who completed the initial telephone interview) were asked to provide blood samples, which were analyzed for pesticide levels and measures of immune competence (e.g., immunoglobulin levels). Residential address matching was performed to compare individuals' serum pesticide levels with distance of their residences from the dump sites. At the time of recruitment into this part of the study, participants were once again asked to provide their physical street address as well as the nearest intersecting street (even if a very small street). Because of the rural nature of the study areas, verbal responses regarding the physical location of residences did not always provide sufficient information to accurately locate them. Therefore, since the participants had to travel to a health care center to have their blood drawn, they were asked at that time to mark the location of their residence on a county planning map (Scale 1 in = 2,000 ft) on which research staff had already encircled their nearest pair of intersecting streets.

A file containing participants' addresses (street and number) was matched, using Arc/Info, against data from the North Carolina Department of Public Instruction's Transportation Information Management System (TIMS) files (which are used for school bus routing and the assignment of bus stops) by a GIS analyst at the North Carolina State Center for Health Statistics. The U.S. TIGER/Line files could not be used because they did not include adequate address information for the streets in the county containing the study areas. The initial automated address-match rate using the TIMS files was 28%. Another 30% were matched with accuracy to the level of the intersecting streets. Maps marked by the study participants allowed the identification of the point locations of the remaining residential addresses as well as more exact locations for the 30% previously digitized by intersection. Geographic coordinates for the dump sites were obtained from the Agency for Toxic Substances and Disease Registry.

The final GIS product for this study was a database from which point locations for both study participant residences and the dump sites could be visually displayed and distances between residences and the dump sites could be calculated. In addition, other information collected from the participants (e.g., demographic information, pesticide levels) could be geographically displayed and analyzed with respect to spatial patterns and distance from the dump sites. To evaluate whether proximity to the dump sites was associated with blood pesticide levels, statistical models could be constructed regressing distance on pesticide levels controlling for potential confounders and/or modifiers of the association such as age, gender, race, length of residence, and occupational or other exposures to pesticides. Residential addresses could also be linked to a growing number of existing computerized geographic databases containing, for example, groundwater flow information.

Technical issues. The study described above suggests potential uses of a GIS and illustrates common problems with address geocoding, especially in rural areas. The initial automated address-match rate of 28% was not unusual for a rural area. Staff of the Carolina Population Center in North Carolina, who have done GIS address matching for a number of projects within the state, report a range of automated address-match rates from a low of approximately 20% in very rural counties to a high of approximately 98% in the largest urbanized county (Philip H. Page, Carolina Population Center, personal communication).

Factors hindering address matching include incomplete or inaccurate information in the street network files (e.g. TIGER/Line and TIMS) used for address matching, lack of standardization of street addresses (e.g., E. Main Street vs. Main Street, East), and lack of assignment of numerical street addresses, especially in rural areas. Steps that can be taken to improve address-match rates include 1) regular updating of street network files to include newly added, changed, or renamed streets (some commercial address geocoding service providers enhance their geographic street databases beyond those found in the original U.S. TIGER/Line files and sell them to people who want to do their own address-matching); 2) standardization of street addresses, although specific standardization rules may vary from one GIS to another; and 3) adoption of city-style addresses in rural areas (fortunately, this is happening in many areas, often due to the establishment of 911 emergency systems in which phone numbers are linked to geographic street addresses).

Ethical issues. Traditionally, the confidentiality of health records used in mapping and spatial applications has been maintained by aggregating data to geographic units (e.g., counties) that are large enough to eliminate the risk of disclosing information at the individual level. The advent of GIS technology has allowed researchers to display the precise location of individual residences. The presentation of residential locations on maps may violate confidentiality, especially if the study area is small or if the number of events per population is low. The challenge to the researcher is to protect the confidentiality of the individual while maintaining locational integrity for spatial analyses.

Since address information is considered a personal identifier, some agencies will not release such data to researchers and some edit the data for reasons of confidentiality. To protect information about individuals in small geographic units, the U.S. Bureau of the Census uses a technique called confidentiality edit. Selected responses to questions among a random sample of households are switched with responses from similar households in the same state.

While mainstream GIS software has not yet developed procedures to protect confidentiality, some researchers have found their own solutions. Some have released small-scale maps or displayed only aggregate data, while others have displaced points to conceal their true locations. Thomas et al. (*20*) coded gonorrhea cases only to their block group area identifier. In their presentation of the data, they divided all of the block groups into quartiles of cases/population and shaded the block groups from highest to lowest prevalence of gonorrhea. Rushton et al. (*21*) used a random displacement algorithm to reveal the pattern but distort the actual location of infant deaths in Des Moines, Iowa. This was accomplished by using the random number generation capability of Microsoft Excel 5.0 (Microsoft, Redmond, WA) and incorporating the random numbers into a GIS to produce altered latitude/longitude coordinates (Gerard Rushton, personal communication). Clearly, confidentiality is an issue that needs to be considered in the presentation of study results.

Epidemiologic issues. With regard to the use of address geocoding for exposure assessment, residential location or any one location may not be the relevant site of exposure. For example, occupational location may be as important or more important than residential location. Therefore, one might need to code a person to his/her occupational location or take into consideration the amount of time the person spends in different locations.

Furthermore, distance from a residence to a source of contamination may not be synonymous with exposure. For example, wind direction or groundwater flow may influence exposure levels. With GIS technology, one can estimate exposures within a geographic region two ways: 1) through spatial interpolation of measured data points, where the levels of exposure between measured data points are estimated (for example, air pollution levels estimated from levels measured at various monitoring stations) or 2) through modeling techniques. Assuming appropriate measured data points exist, several spatial interpolation techniques, available on most GISs, can be used to estimate exposures across a region (*3*). If measured data points do not exist, then one can estimate exposure levels through the modeling of information related to exposure dispersion. For example, if one were interested in estimating pesticide levels across a region, one might want to include in the model the locations of the crops on which the pesticides are used, groundwater flow, soil type, and leaching potential of the pesticides. Models tend to have greater predictive value when 1) the data used to create them are accurate, 2) the conditions under which they are used are relatively simple, and 3) the geographic area to which they apply is close to the source(s) of exposure (*3*).

Hardware and Software Requirements

Geographic information systems have been developed for a variety of computer environments from high-powered workstations to low-end personal computers (PC). Workstation GISs are usually Unix-based and have larger data storage and processing capabilities. Arc/Info is an example of a powerful workstation GIS product; a PC version also exists. Arc/Info has a several-year learning curve and is mostly used by professionals as it requires knowledge of geographic concepts and spatial analysis techniques.

Among the recent developments in GIS software there has been a transition from the command line interface (CLI) software, such as Arc/Info, to the more user-friendly PC-based graphical user interface (GUI) software (*22*). Several of these software products were recently reviewed by Thrall et al. (*23*) and include ArcView (ESRI), Maptitude (Caliper Corporation, Newton, MA), MapInfo (MapInfo Corp., Troy, NY), and AtlasGIS (formerly, Strategic Mapping, Inc., now, ESRI). These products have a shorter learning curve than the CLI software and include menu interfaces and tool bars that can be used in a point and click environment. All of these products allow users to easily produce shaded maps and provide users with a range of data classification and map symbol options. They also include basic GIS functions such as address matching, spatial query, distance calculations, and point in polygon analysis (determining which polygon a specific point falls in and assigning attributes from that polygon to the point in question). Script languages are available and provide opportunities for greater customization of menus and applications. The spatial analysis capabilities of these products vary, but in general they are not as powerful as the CLI products. Some, such as ArcView, have add-on modules for enhanced spatial analysis capabilities.

EpiMap (USD, Inc., Stone Mountain, GA), a mapping software developed for IBM-compatible microcomputers, is much more limited in its capabilities than the products previously described. It creates maps from predetermined geographic boundary files (e.g., county or a single state) and can incorporate nongeographic data from EpiInfo (USD, Inc.) a dBase file (Borland International, Inc., Scotts Valley, CA), or direct keyboard entry.

Selection of GIS software should be based on the needs of the researcher and the capabilities of the software. Some questions to consider when selecting a GIS software product include: How easy is the software to learn and use? Is there good documentation and technical support? Is the company stable and what are its long range plans for future development of the product? Will the features of the software meet the demands of the current research project? How about future projects? While some GIS applications, such as county level mapping of disease rates with existing databases, require little special training, it is important that all GIS users have some knowledge of geographic concepts and principles of map design. For complicated or customized analysis of spatial data, consultation with a GIS specialist is recommended.

Economic considerations. Unless a GIS is going to be used frequently and/or on a large scale, the investment in time to become familiar with the software and money for software and equipment may be prohibitive. As pointed out in the example in rural North Carolina, some GIS functions can be done manually. However, in knowledgeable hands, a GIS can significantly reduce the time necessary to accomplish certain tasks (e.g., address matching, and the calculation of distance from residences to a source of exposure) and can open up new and more complex avenues for the display and analysis of exposure/disease associations. Collaboration with a GIS specialist or hiring the services of a commercial vendor to perform some of the GIS functions, such as address matching, may be more cost effective than for the epidemiologist to invest time and money in becoming proficient in the use of GIS technology.

Conclusions and Recommendations

GIS technology is a tool of great potential for environmental health researchers. It can be used to support or suggest hypotheses regarding disease causation through the conduct of relatively quick and inexpensive ecologic studies using existing databases and easily computerized data. For example, variations in disease rates across geographic regions and differences in disease rates within and across regions over time can be screened with GIS methods. Because a GIS can easily manipulate large amounts of data, it can facilitate analyses on the local, regional, or national level. Environmental exposures that occur within noncontiguous regions can be aggregated to enhance study population sizes.

GIS technology can also be used to test hypotheses regarding environmental risk factors for disease in individual level studies (cross-sectional, case–control, and cohort studies), which have advantages over ecologic studies, especially with regard to their ability to control for potential confounders and modifiers of exposure/disease associations. In these studies, the outcome or disease data (and often the confounder information) are obtained on the individual level, while the GIS-derived exposure information is more ecologic in nature, often pertaining to a geographic region. Data from multiple sources can be integrated and modeled to estimate exposures.

Finally, GIS technology can also be used to simplify or expedite certain steps necessary to conduct epidemiologic research. Examples presented earlier indicate how a GIS can be used to select geographic regions with specific characteristics for inclusion in a study, to identify geographically eligible study participants, and to calculate residential proximity to exposure sources through address geocoding.

The degree of confidence one has in the results of environmental epidemiologic investigations depends to a large extent on the accuracy of the exposure information. Concerns regarding GIS-derived exposure assessments relate to 1) the aggregate/ecologic nature of the data, which may not be relevant at the individual level; 2) the quality of the data that are input into the GIS (e.g., accuracy and completeness), 3) the appropriateness of combining multiple databases; and 4) the relevance of the map layers to the exposure/outcome association of interest (e.g., timing of the data collection). While GIS technology may enhance epidemiologic research by making some steps quicker, easier, and cheaper to accomplish, it will not replace traditional epidemiologic methods and approaches.

Collaboration and communication among researchers from a variety of fields including epidemiology, medical geography, environmental sciences, and biostatistics are necessary to realize the full potential of GIS technology in environmental health studies. Whereas epidemiologists are well-versed in study design issues related to the collection and analysis of data on individuals, medical geographers are knowledgeable about the integration of group level data from many sources and are the experts on spatial study design and analysis (*24,25*). Environmental scientists are needed to help determine the appropriate information to include in GIS exposure estimation models, and the expertise of biostatisticians is essential to develop, perform, and interpret sophisticated spatial statistical analyses. For example, collaborative efforts proved beneficial in a multidisciplinary study of infant deaths in Des Moines, Iowa (*21*). When a pediatrician suggested mapping infant deaths by census tract, the medical geographer noted that ignoring predetermined political boundaries would provide a better means of revealing the spatial distribution of infant deaths in the city.

Measures to enhance communication among researchers in different fields include 1) publication of health studies incorporating GIS technology in journals read by epidemiologists, environmental scientists, and biostatisticians; 2) inclusion of GIS as a subject heading on MEDLINE; 3) attendance by environmental health researchers at professional meetings devoted to the discussion of the use of GIS technology in health research; and 4) allocation of research funds that encourage collaborative efforts among researchers from different fields.

The full potential of GIS technology in health research is not yet known. Collaborative efforts could lead to creative and practical applications of GIS technology to answer complex environmental epidemiologic research questions.

References

1. Antenucci JC, Brown K, Croswell PL, Kevany MJ. Geographic Information Systems: A Guide to the Technology. New York:Van Nostrand Reinhold, 1991.
2. Goodchild MF. The state of GIS for environmental problem-solving. In: Environmental Modelling with GIS (Goodchild MF, Parks BO, Steyaert LT, eds). New York:Oxford University Press, 1993;8–15.
3. Briggs DJ, Elliot P. The use of geographical information systems in studies on environment and health. World Health Stat Q 48:85–94 (1995).
4. Frisch JD, Shaw GM, Harris JA. Epidemiologic research using existing databases of environmental measures. Arch Environ Health 45:303–307 (1990).
5. Twigg L. Health based geographical information systems: their potential examined in the light of existing data sources. Soc Sci Med 30:143–155 (1990).
6. Shrestha RL, Dewitt BA, Wilson MM. Consideration and effect of a local base station range and horizontal position determination by GPS techniques. Sur Land Inf Syst 55:39–49 (1995).
7. Stallones L, Nuckols JR, Berry JK. Surveillance around hazardous waste sites: geographic information systems and reproductive outcomes. Environ Res 59:81–92 (1992).
8. Waller LA. Geographic information systems and environmental health. Health Environ Digest 10:85–88 (1996).
9. Greenland S, Robins J. Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. Am J Epidemiol 139:747–760 (1994).
10. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 18:269–274 (1989).
11. Morgenstern H. Ecologic studies in epidemiology: concepts, principles, and methods. Annu Rev Public Health 16:61–81 (1995).
12. Hertz-Picciotto I. Environmental Epidemiology. In: Modern Epidemiology (Rothman KJ, Greenland S, eds). 2nd ed. Boston:Little, Brown and Company. In press.
13. Odland J. Spatial autocorrelation. In: Spatial Autocorrelation. Scientific Geography Series, vol 9 (Thrall GI, ed). London:Sage Publications, 1988;7–17.
14. Cressie NAC. Statistics for Spatial Data. New York:John Wiley and Sons, 1993.
15. Lawson A, Waller L, Biggeri A, eds. Special issue: spatial disease patterns. Stat Med 14(21/22): 2289–2508 (1995).
16. Guthe WG, Tucker RK, Murphy EA, England R, Stevenson E, Luckhardt JC. Reassessment of lead exposure in New Jersey using GIS technology. Environ Res 59:318–325 (1992).
17. Brody DJ, Pirkle JL, Kramer RA, Flegal KM, Matte TD, Gunter EW, Paschal DC. Blood lead levels in the US population: phase I of the Third National Health and Nutrition Examination Survey (NHANES III, 1988–1991. JAMA 272: 277–283 (1994).
18. Wartenberg D, Greenberg M, Lathrop R. Identification and characterization of populations living near high-voltage transmission lines: a pilot study. Environ Health Perspect 101:626–632 (1993).
19. Glass GE, Schwartz BS, Morgan JM III, Johnson DT, Noy PM, Israel E. Environmental risk factors for Lyme disease identified with geographic information systems. Am J Public Health 85:944–948 (1995).
20. Thomas JC, Schoenbach VJ, Weiner DH, Parker EA, Earp JA. Rural gonorrhea in the southeastern United States: a neglected epidemic? Am J Epidemiol 143:269–277 (1996).
21. Rushton G, Krishnamurti D, Krishnamurthy R, Song H. A geographic information analysis of urban infant mortality rates. Geo Info Systems 5:52–56 (1995).
22. Thrall IG. New generation of mass-market GIS software: a commentary. Geo Info Systems 5:58–60 (1995).
23. Thrall IG, del Valle J, Elshaw-Thrall S. First impressions—four mass market GIS software programs. Geo Info Systems 5:60–65 (1995).
24. Cliff AD, Haggett P. Atlas of Disease Distributions: Analytic Approaches to Epidemiological Data. Oxford:Basil Blackwell, 1988.
25. Meade M, Florin J, Gesler W. Medical Geography. New York:Guilford, 1988.